Assignment 06

In this assignment, you use the insurance data set.

The main code is insurance.m

The computeCost function takes the data set X, for each example/observation in X, does the following (Here L = 3 indicating 3 layers):

a. First the labels are converted:

"1", which is "No" is converted to a vector of [1,0]; and "2", which is "Yes" is converted to [0,1] using the identity matrix trick. The new labels are stored in the variable Y.

b. Forward propagation**:** For each example/observation x

- $a^{(1)} = x$
- Add bias value ${a^{(1)}}_0 = 1$
- For l =2… L
  - $z^{(l)} = \theta^{(l-1)} a^{(l-1)}$
  - $a^{(l)} = g(z^{(l)})$
  - Add bias value ${a^{(l)}}_0 = 1$
- End
- $h_\theta(x) = a^{(L)}$

c. Now compute the regularized cost function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\left(-{y^{(i)}}_k log h_\theta\left(x^{(i)}\right)_k - \left(1-{y^{(i)}}_k\right) log\left(1-h_\theta(x^{(i)})_k\right)\right)$$

$$+\frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left({\theta^{(l)}}_{ij}\right)^2$$

d. Now the function starts back-propagating

- $\delta^{(L)} = a^{(L)} - y$ (now for all output nodes, $\delta^{(L)}$ is a vector)

- For l=L-1…2
    - $\delta^{(l)}=a^{(l)}\left(1-a^{(l)}\right).*\theta^{(l)^T}\delta^{(l+1)}$
- End

e. And compute the gradient of the cost function

- $\Delta^{(l)}$ = 0, l=1..L-1
- For each instance in the training dataset
    - Forward propagation to compute $a^{(l)}$, l=1..L
    - Backpropagation to compute $\delta^{(l)}$, l=L..2
    - Compute $\Delta^{(l)}=\Delta^{(l)}+\delta^{(l+1)^T}a^{(l)}$, l=1..L-1
- End
- Compute $\frac{\partial J(\theta)}{\partial\theta^{(l)}}=\frac{1}{m}\Delta^{(l)}+\lambda\theta^{(l)}$

The cost and the gradient of the cost are then stored in the variables J and grad.

A1: contains multiple $a^{(1)}$ on its rows (after adding the bias value), each is corresponding to an example/observation

A2: contains multiple $a^{(2)}$ on its rows (after adding the bias value), each is corresponding to an example/observation

A3: contains multiple $a^{(3)}$ on its rows, each is corresponding to an example/observation. Note that A3 is also H.

Z2: contains multiple $z^{(2)}$ on its rows, each is corresponding to an example/observation.

Z3: contains multiple $z^{(3)}$ on its rows, each is corresponding to an example/observation.

delta3, delta2, BigDelta1, BigDelta2 are matrix implementation of the corresponding variables in the pseudo-code.

The normalize function is used to normalize the features/variables in X. The predict function is to predict if a customer would buy insurance given X. The prediction is stored in pred and returned to the main code. Others include sigmoid, and gradient.

1. (4 points) Read, understand, and run the MATLAB code.

Pick either R or Python to do the work based on the working MATLAB code is given to you. 1 additional point if you write code in both Python, and R. Logics, variable names, and function names should follow the ones in the given MATLAB code as much as possible.

In Python, you can use the minimize function in the scipy.optimize package:

http://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html

There are several optimizations methods by the method argument. You can use any, or can use your own Gradient Decent.

In R, use the optim function in the optim package

https://stat.ethz.ch/R-manual/R-devel/library/stats/html/optim.html

However, there is no restriction on what package you can use

2. (1 point) Explain your code (you can do this line by line, or small blocks of code).

Submission: your script, and a document explaining your work.